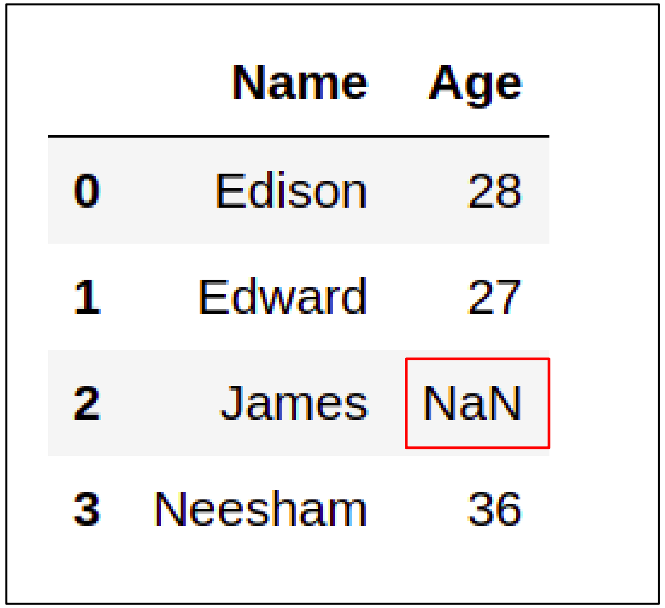
Handling Missing Data

Dealing with Missing values (1/6)

- Data often contains missing values.
- A missing value means loss of information.
- In pandas, an NaN indicates a missing value.
- Consider the following dataframe called data with one missing value in the ‘Age’ column.

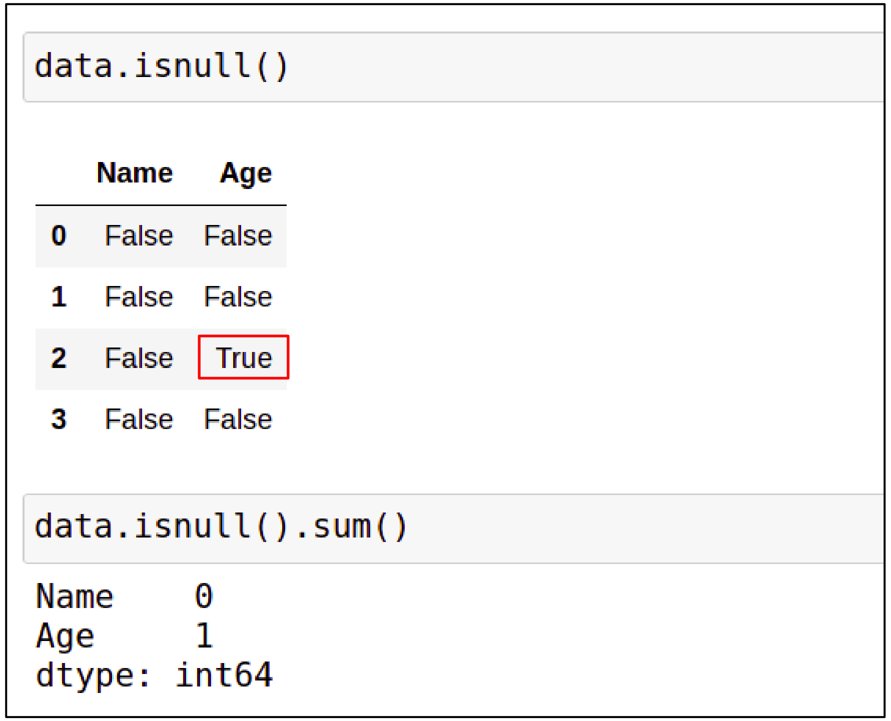


	Name	Age
0	Edison	28
1	Edward	27
2	James	NaN
3	Neesham	36

Dealing with Missing values (2/6)

Finding Missing Values in Pandas

- The .isnull() function tells us if a cell is empty or not.
- Use the .sum() function with the .isnull() function to find total number of missing values in the data.



```
data.isnull()
```

	Name	Age
0	False	False
1	False	False
2	False	True
3	False	False

```
data.isnull().sum()
```

```
Name     0
Age      1
dtype: int64
```

Dealing with Missing values (3/6)

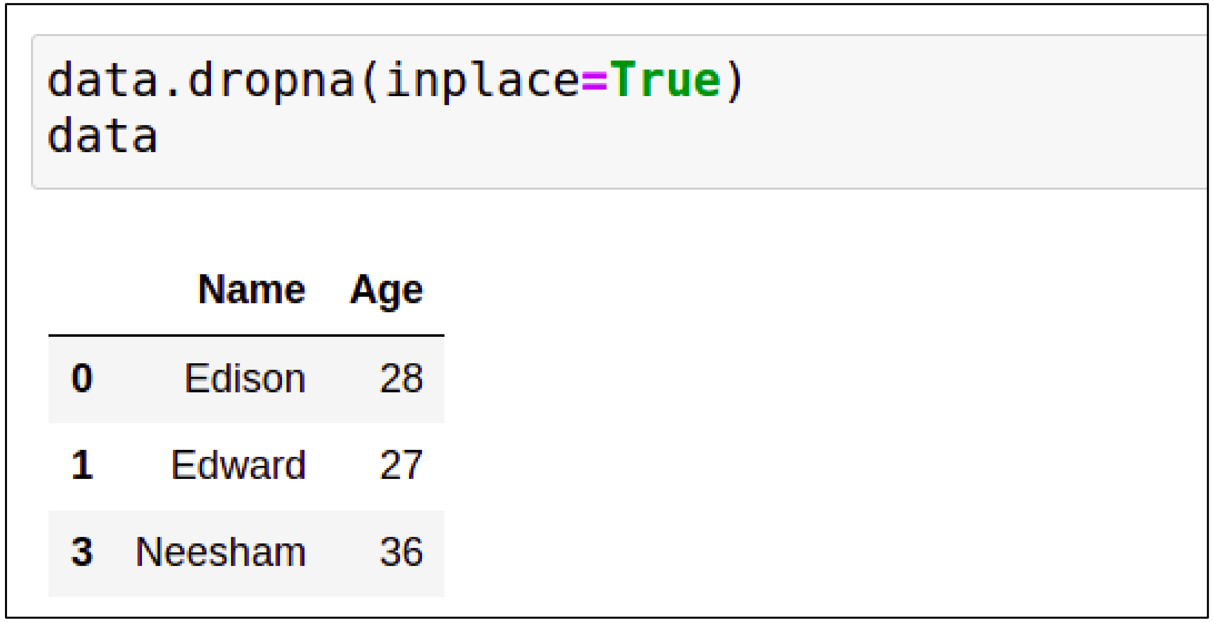
Dealing with Missing Values in Pandas

- There are a number of ways to deal with these missing values.
- Which method to use depends upon the kind of data and the task that the data is supposed to accomplish.
- Different Methods Used are;
  - Deleting rows with missing values.
  - Replacing missing values with mean/median/mode.

Dealing with Missing values (4/6)

Deleting Rows with Missing Values

- One way to deal with the missing values is to delete the rows containing missing values.
- Use the .dropna() with inplace set to True to remove missing values from the dataset.



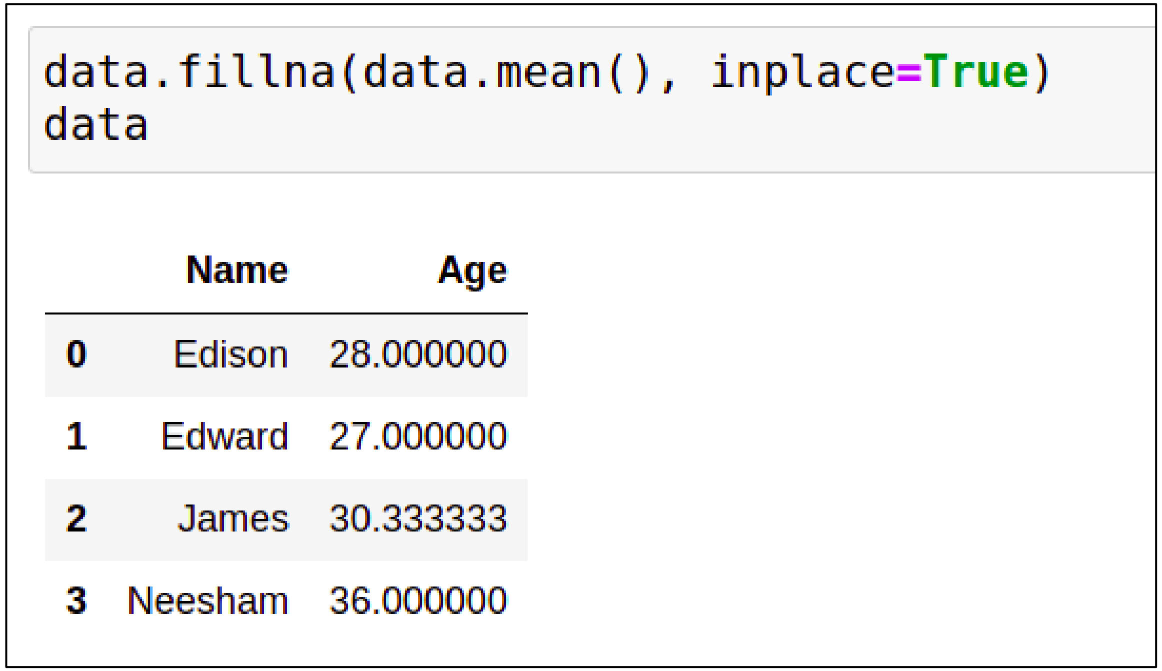
```
data.dropna(inplace=True)
data
```

	Name	Age
0	Edison	28
1	Edward	27
3	Neesham	36

Dealing with Missing values (5/6)

Replacing Missing Values with Mean/Median/Mode

- We can also replace the missing values in each column with one of the statistical measures (mean/median/mode) of that column.
- Use the .fillna() method to fill the missing values with mean, median, or mode.



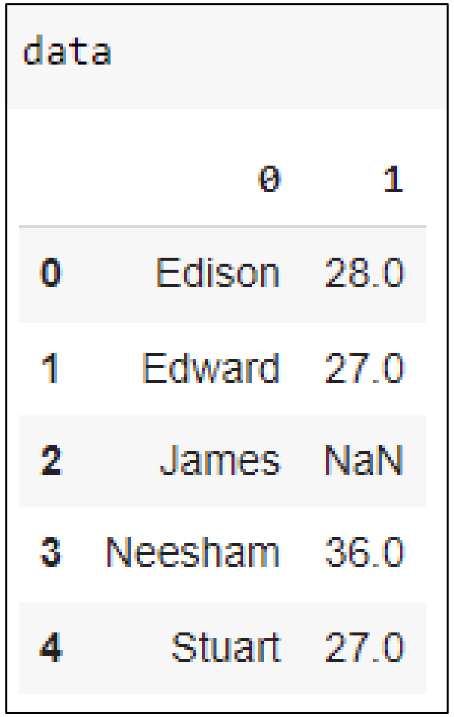
```python
data.fillna(data.mean(), inplace=True)
data
```

	Name	Age
0	Edison	28.000000
1	Edward	27.000000
2	James	30.333333
3	Neesham	36.000000

Dealing with Missing values (6/6)

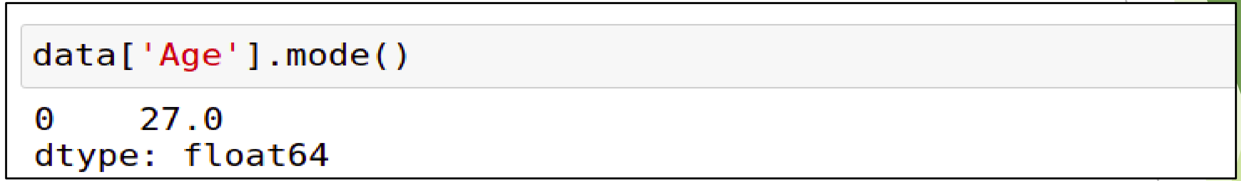
Replacing Missing Values with Mean/Median/Mode

- In this example, we replace the missing value in the ‘Age’ column with the mode of the ‘Age’ column.



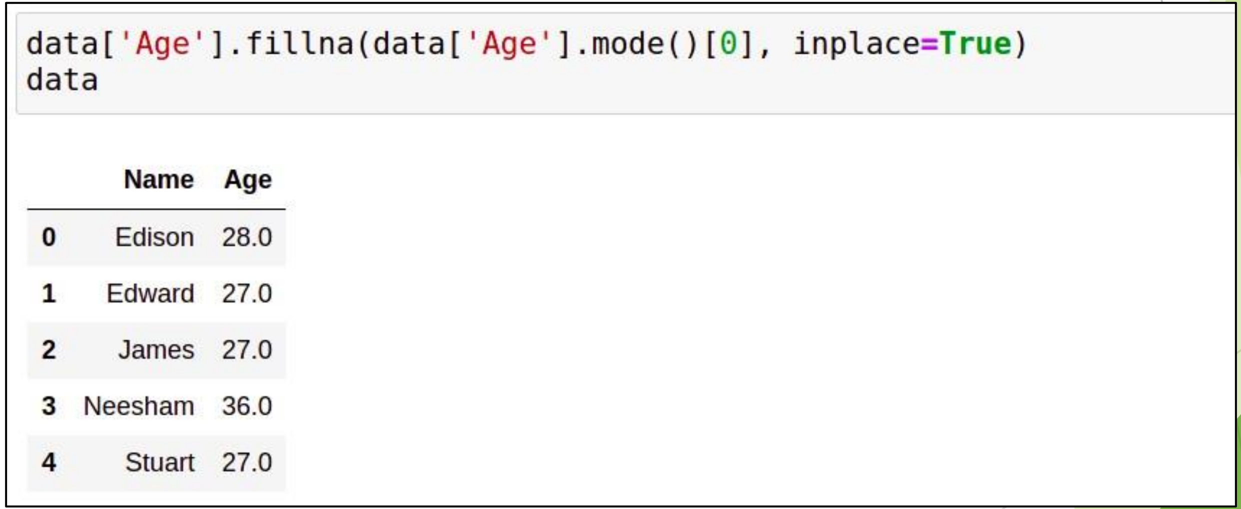
```
data
```

	0	1
0	Edison	28.0
1	Edward	27.0
2	James	NaN
3	Neesham	36.0
4	Stuart	27.0



```
data['Age'].mode()
```

```
0    27.0
dtype: float64
```



```
data['Age'].fillna(data['Age'].mode()[0], inplace=True)
data
```

	Name	Age
0	Edison	28.0
1	Edward	27.0
2	James	27.0
3	Neesham	36.0
4	Stuart	27.0